Butlletí de la Societat Catalana de Matemàtiques
Vol. 21, núm. 1, 2006. Pàg. 39–52

Combinatòria i biologia: funcions d'inferència i alineació de seqüències*

Sergi Elizalde

> La manca de contacte real entre les matemàtiques i la biologia és o bé una tragèdia, o bé un escàndol, o bé un repte; és difícil decidir quin és el cas.
>
> **Gian-Carlo Rota**

Resum Aquest article mostra alguns exemples d'aplicació d'eines combinatòries a problemes en biologia computacional. Els models estadístics s'usen per resoldre qüestions provinents de la biologia, com per exemple per determinar quines parts del genoma es tradueixen a proteïnes, o com una seqüència d'ADN es va transformar en una altra durant l'evolució, a través d'una sèrie de mutacions, insercions i supressions. Cada possible resposta té una certa probabilitat que depèn dels paràmetres del model. Quan aquests es coneixen, la resposta més probable, anomenada *explicació*, s'obté resolent un problema d'optimització combinatòria. La funció que envia cada observació a la seva explicació corresponent s'anomena *funció d'inferència*.

En aquest article donem una fita superior al nombre de funcions d'inferència de qualsevol model gràfic dirigit. Aquesta fita és polinòmica en la mida del model, per un nombre fix de paràmetres, i representa una millora respecte a la fita exponencial de Pachter i Sturmfels que es coneixia fins ara. Després apliquem aquesta fita a un model per alineació de seqüències que s'utilitza en biologia computacional, i veiem que en aquest cas la nostra fita és asimptòticament ajustada.

Paraules clau: funcions d'inferència, models gràfics, alineació de seqüències, politop de Newton.

Classificació MSC2000: 62F15, 52C45, 52B20, 62P10, 52B05.

* Versió 1.0, gener de 2006.

1 Introducció

Donat un conjunt de dades observades, hi ha molts models estadístics que es fan servir per trobar les dades *ocultes* (no observades) que expliquen millor les observacions. En aquest article considerem models gràfics, també anomenats xarxes bayesianes, una classe àmplia que inclou models utilitzats freqüentment, com ara els models de Markov ocults (HMM). Aquests models relacionen les dades observades i les ocultes probabilísticament. Un problema natural consisteix a determinar, donada una observació particular, quines són les dades ocultes més probables (que s'anomenen l'*explicació*). Els paràmetres d'aquests models són les probabilitats que lliguen les dades observades i les ocultes. Qualssevol valors fixats dels paràmetres determinen una manera d'assignar una explicació a cada possible observació. Això ens dóna una funció, anomenada *funció d'inferència*, que envia cada observació a l'explicació corresponent.

En l'article estudiem el problema de determinar el nombre de funcions d'inferència. Aquest nombre és important perquè indica com els paràmetres del model poden afectar la solució. Un exemple d'una funció d'inferència és la popular opció «Did you mean» o «Potser ha volgut dir» de *google*, que es podria implementar amb un model de Markov ocult, on les dades observades són els caràcters que introduïm a l'ordinador, i les dades ocultes són els que en realitat volíem introduir. Els models gràfics s'usen sovint en aquests tipus d'enfocaments a intel·ligència artificial (podeu trobar-ne una introducció a [6]).

Les funcions d'inferència per a models gràfics són també importants en la biologia computacional [7, secció 1.5]. Un bon exemple són els models d'alineació òptima de seqüències [7, secció 2.2]. En aquest cas, les funcions d'inferència indiquen, donat un parell de seqüències d'ADN, quin és el procés d'evolució més probable que va seguir l'una per transformar-se en l'altra. Més precisament, una funció d'inferència envia cada parella de seqüències d'ADN a una alineació òptima de les seqüències. Si canviem els paràmetres del model, és possible que variïn les alineacions òptimes, i, per tant, les funcions d'inferència poden canviar.

Una altra classe d'exemples que trobem la biologia són les anomenades *funcions cercadores de gens*, que apareixen a [8, secció 5]. Aquestes funcions d'inferència (corresponents a un HMM concret) s'utilitzen per identificar estructures de gens en seqüències d'ADN. Les funcions cercadores de gens determinen quines parts d'una seqüència donada d'ADN són *exons* i quines parts són *introns*. Aquesta distinció és important perquè els exons són les parts que codifiquen proteïnes, mentre que els introns, que formen la major part del genoma, són segments intermedis la funció dels quals no és coneguda perquè no són mai traduïts a proteïnes. Una observació en aquest model és una seqüència de nucleòtids en l'alfabet $\Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, i una explicació és una seqüència d'uns i zeros que indica si cada nucleòtid concret és en un gen o no. L'objectiu és fer servir la informació en les observacions (que es pot trobar a través de la seqüenciació d'ADN) per decidir quina és la informació oculta que indica quins nucleòtids formen part de gens (cosa difícil d'esbrinar directament).

Una conclusió a la qual arribarem en aquest article és que no hi pot haver gaires funcions d'inferència diferents, encara que els paràmetres puguin variar contínuament entre tots els valors possibles. Per exemple, en el HMM binari homogeni de longitud 5 (en trobareu les definicions a la secció 2.1), l'observació és una seqüència binària de longitud 5, i l'explicació també serà una seqüència binària de longitud 5. A primera vista, hi ha

$$32^{32}=1\ 461\ 501\ 637\ 330\ 902\ 918\ 203\ 684\ 832\ 716\ 283\ 019\ 655\ 932\ 542\ 976$$

possibles funcions de seqüències observades a explicacions. Però resulta que només 5.266 de totes aquestes possibles funcions són, de fet, funcions d'inferència.

Diferents funcions d'inferència representen diferents criteris per decidir quina és l'explicació més probable per a cada observació. Una fita en el nombre de funcions d'inferència és important perquè indica com un model pot respondre a canvis en els valors dels paràmetres (dels quals normalment només es té un coneixement aproximat). D'altra banda, la fita polinòmica que donem a la secció 3 suggereix que podria ser factible precomputar totes les funcions d'inferència d'un model gràfic donat, la qual cosa donaria una manera eficient de proporcionar una explicació per a cada observació donada.

L'estructura d'aquest article és la següent. Comencem introduint alguns preliminars sobre models gràfics i funcions d'inferència a la secció 2, i també alguns fets sobre politops. A la secció 3 presentem el resultat principal, que diu que, en qualsevol model gràfic, el nombre de funcions d'inferència creix polinomialment en la mida del model (si fixem el nombre de paràmetres). A la secció 4 apliquem el teorema a un model per alineació de seqüències, i provem que la fita és asimptòticament exacta en aquest model. En aquest cas, la fita és quadràtica en la mida de les seqüències d'ADN que volem alinear.

Tot i que la majoria d'aquests problemes provenen de situacions reals en biologia i en genètica, poden ser fàcilment traduïts a qüestions purament matemàtiques, cosa que permet estudiar-los fent servir eines de la combinatòria. Aquest és un camp on es barregen l'àlgebra, la combinatòria, l'estadística, la biologia i la computació.

2 Preliminars

2.1 Models gràfics

Un *model gràfic* és una família de distribucions de probabilitat conjuntes per a una col·lecció de variables aleatòries discretes $\mathbf{Z} = (Z_1, \ldots, Z_m)$, on cada Z_i pren valors en un espai d'estats finit Σ_i. Aquí ens concentrarem en models gràfics dirigits. Un *model gràfic dirigit* (o *xarxa bayesiana*) és un graf dirigit (digraf) finit i acíclic G on cada vèrtex u_i correspon a una variable aleatòria Z_i.

Cada vèrtex u_i també té una funció de probabilitat associada

$$p_i : \left(\prod_{j:\ u_j \text{ és un pare de } u_i} \Sigma_j \right) \to [0,1]^{|\Sigma_i|},$$

on diem que u_j és pare de u_i si hi ha una aresta de u_j cap a u_i. Donats els estats de tots els Z_j pels quals u_j és un pare de u_i, la probabilitat que u_i estigui en un estat donat és independent dels valors de tots els altres vèrtexs que no són descendents de u_i, i la funció p_i dóna aquesta probabilitat. En particular, tenim l'igualtat

$$\begin{aligned} \text{Prob}(\mathbf{Z} = \tau) &= \prod_i \text{Prob}\left(Z_i = \tau_i, \text{ condicionat a } Z_j = \tau_j \text{ per cada pare } u_j \text{ de } u_i\right) \\ &= \prod_i \left[p_i\left(\tau_{j_1}, \ldots, \tau_{j_r}\right) \right]_{\tau_i}, \end{aligned}$$

on $u_{j_i}, \ldots, u_{j_r}$ són els pares de u_i. Normalment es considera que en els vèrtexs sense pares hi ha la distribució de probabilitat uniforme en els seus estats, tot i que distribucions de probabilitat més generals també són possibles. Consulteu [7, secció 1.5] per a una introducció als models gràfics.

EXEMPLE El model de Markov ocult (HMM) és un model amb variables aleatòries $\mathbf{X} = (X_1, \ldots, X_n)$ i $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Les arestes van de X_i a X_{i+1} i de X_i a Y_i.

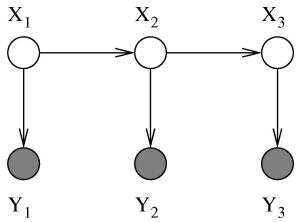

FIGURA 1: El graf d'un HMM per $n = 3$.

En general, cada X_i té el mateix espai d'estats, Σ, i cada Y_i té el mateix espai d'estats, Σ'. Un HMM s'anomena *homogeni* si les funcions de probabilitat p_{X_i}, per $1 \le i \le n$, són idèntiques entre si i les p_{Y_i} també. En aquest cas, cadascuna de les p_{X_i} és especificada per una mateixa matriu $T = \left(t_{jk}\right)$ de dimensions $|\Sigma| \times |\Sigma|$ (la matriu de *transició*), on t_{jk} és la probabilitat que X_{i+1} prengui el valor τ_k si X_i pren el valor τ_j. Anàlogament, cadascuna de les p_{Y_i} correspon a una mateixa matriu $S = \left(s_{jk}\right)$, de dimensions $|\Sigma| \times |\Sigma'|$ (la matriu d'*emissió*).

A l'exemple hem separat les variables en dos conjunts. En models gràfics generals també tenim dos tipus de variables: variables observades $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ i variables ocultes $\mathbf{X} = (X_1, X_2, \ldots, X_q)$. Normalment les

variables observades són els vèrtexs sense descendents, però no hi ha cap raó per ser sempre així. Per simplificar la notació, podem assumir, com sol ser el cas a la pràctica, que totes les variables observades prenen valors en el mateix alfabet finit, Σ', i que totes les variables ocultes prenen valors en el mateix alfabet finit, Σ.

Observeu que, donats Σ i Σ', els HMM d'aquest exemple depenen només d'un conjunt fix de paràmetres, t_{jk} i s_{jk}, encara que n creixi. Aquests són els tipus de models que ens interessen. Fixarem un enter d que serà el nombre de paràmetres, i anomenarem els paràmetres $\theta_1, \theta_2, \ldots, \theta_d$. Un model gràfic amb d paràmetres voldrà dir un model gràfic on cada probabilitat $\left[p_i\left(\tau_{j_1}, \ldots, \tau_{j_r}\right)\right]_{\tau_k}$ és un monomi en els paràmetres, i a més el grau d'aquest monomi està fitat pel nombre de pares de u_i. Per exemple, en el HMM homogeni, cada u_i té només un pare, i les coordenades de p_i són monomis de grau 1 (o bé t_{jk} o bé s_{jk}).

D'aquí endavant E denotarà el nombre d'arestes del graf subjacent d'un model gràfic, n serà el nombre de variables aleatòries observades, i q, el nombre de variables aleatòries ocultes. Les observacions, doncs, són seqüències de $(\Sigma')^n$ i les explicacions són seqüències de Σ^q. Definim $l = |\Sigma|$ i $l' = |\Sigma'|$.

Per a cada observació τ i informació oculta $\mathbf{h}$, Prob $(\mathbf{X} = \mathbf{h},\ \mathbf{Y} = \tau)$ és un monomi $f_{\mathbf{h},\tau}$ de grau com a molt E en els paràmetres $\theta_1, \theta_2, \ldots, \theta_d$. Per a cada observació $\tau \in (\Sigma')^n$, la probabilitat Prob$(\mathbf{Y} = \tau)$ és la suma sobre totes les possibles dades ocultes $\mathbf{h}$ de Prob $(\mathbf{X} = \mathbf{h},\ \mathbf{Y} = \tau)$, i per tant Prob$(\mathbf{Y} = \tau)$ és el polinomi $f_\tau = \sum_{\mathbf{h}} f_{\mathbf{h},\tau}$ en els paràmetres $\theta_1, \theta_2, \ldots, \theta_d$. El grau de f_τ és com a màxim E.

2.2 Funcions d'inferència

Per a valors fixos dels paràmetres, el problema bàsic d'inferència consisteix a determinar, per a cada observació donada τ, el valor $\mathbf{h} \in \Sigma^q$ de les dades ocultes que maximitza Prob$(\mathbf{X} = \mathbf{h} \mid \mathbf{Y} = \tau)$. Una solució d'aquest problema d'optimització es denota $\widehat{\mathbf{h}}$ i s'anomena una *explicació* de l'observació τ. Cada elecció dels valors dels paràmetres $(\theta_1, \theta_2, \ldots, \theta_d)$ defineix una *funció d'inferència* $\tau \mapsto \widehat{\mathbf{h}}$ del conjunt d'observacions $(\Sigma')^n$ al conjunt d'explicacions Σ^q.

És possible que hi hagi més d'un valor de $\widehat{\mathbf{h}}$ que assoleixi el màxim de Prob$(\mathbf{X} = \mathbf{h} \mid \mathbf{Y} = \tau)$. En aquest cas, per simplificar, triarem només una explicació, segons alguna regla consistent per a desfer empats, decidida amb antelació. Una altra alternativa seria definir funcions d'inferència com a funcions de $(\Sigma')^n$ a subconjunts de Σ^q. Això no afectaria els resultats de l'article; així que, per simplificar, considerem només funcions d'inferència com les que hem definit abans.

És interessant observar que el nombre total d'aplicacions $(\Sigma')^n \longrightarrow \Sigma^q$ és $(l^q)^{(l')^n} = l^{q(l')^n}$, que és doblement exponencial en la longitud n de les observacions. Tanmateix, la majoria d'aquestes aplicacions no són funcions d'inferència, per a cap valor dels paràmetres. Fins ara, la millor fita superior coneguda pel nombre de funcions d'inferència era una fita exponencial en la

longitud de les observacions donada a [9, corol·lari 10]. A la secció 3 donarem una fita superior polinòmica al nombre de funcions d'inferència d'un model gràfic.

2.3 Politops

Aquí repassarem alguns fets sobre politops convexos i introduirem la notació. Recordem que un politop és un conjunt fitat obtingut com la intersecció d'un nombre finit de semiespais tancats o, equivalentment, l'envolupant convexa d'un conjunt finit de punts. Per a les definicions bàsiques podeu veure [10].

Donat un polinomi $f(\theta) = \sum_{i=1}^{N} \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}}$, el seu *politop de Newton*, que es denota per $\mathrm{NP}(f)$, es defineix com l'envolupant convexa a $\mathbb{R}^d$ del conjunt de punts $\{(a_{1,i}, a_{2,i}, \ldots, a_{d,i}) : i = 1, \ldots, N\}$. Per exemple, si $f(\theta_1, \theta_2) = 2\theta_1^3 + 3\theta_1^2\theta_2^2 + \theta_1\theta_2^2 + 3\theta_1 + 5\theta_2^4$, llavors el seu politop de Newton $\mathrm{NP}(f)$ és el de l'esquerra de la figura 2.

Donats un politop $P \subset \mathbb{R}^d$ i un vector $w \in \mathbb{R}^d$, el conjunt de tots els punts de P en què el funcional lineal $x \mapsto x \cdot w$ assoleix el màxim determina una *cara* de P. Es denota així:

$$\mathrm{cara}_w(P) \quad = \quad \{\, x \in P \,:\, x \cdot w \geq y \cdot w \ \text{ per tot } y \in P \,\}.$$

Les cares de dimensió 0 (que consisteixen en un sol punt) s'anomenen *vèrtexs* i les cares de dimensió 1 s'anomenen *arestes*. Si d és la dimensió del politop, les cares de dimensió $d-1$ s'anomenen *facetes*.

Sigui P un politop i F una cara de P. El *con normal* de P a F és

$$N_P(F) \quad = \quad \{w \in \mathbb{R}^d \,:\, \mathrm{cara}_w(P) = F\,\}.$$

El conjunt de tots els cons $N_P(F)$ on F és una cara de P es denota per $\mathcal{N}(P)$ i s'anomena el *ventall normal* de P. El ventall normal $\mathcal{N}(P)$ és una partició de $\mathbb{R}^d$ en cons. Els cons de $\mathcal{N}(P)$ estan en bijecció amb les cares de P, i per $w \in N_P(F)$, el funcional lineal $x \cdot w$ assoleix el màxim a F. La figura 2 mostra el ventall normal d'un politop.

La *suma de Minkowski* de dos politops P and P' es defineix com a

$$P + P' := \{\mathbf{x} + \mathbf{x}' : \mathbf{x} \in P, \mathbf{x}' \in P'\}.$$

El *refinament comú* de dos o més ventalls normals és la col·lecció de cons obtinguda com la intersecció d'un con de cada un dels ventalls. Per als politops $P_1, P_2, \ldots, P_k$, el refinament comú dels seus ventalls normals es denota $\mathcal{N}(P_1) \wedge \cdots \wedge \mathcal{N}(P_k)$. El lema següent ens diu que el ventall normal d'una suma de Minkowski de politops és el refinament comú dels seus cons individuals (vegeu [10, proposició 7.12] o bé [3, lema 2.1.5]):

1 Lema $\mathcal{N}(P_1 + \cdots + P_k) = \mathcal{N}(P_1) \wedge \cdots \wedge \mathcal{N}(P_k)$.

Acabem amb un resultat de Gritzmann i Sturmfels que serà útil més endavant i que dóna una fita al nombre de vèrtexs d'una suma de Minkowski de politops.

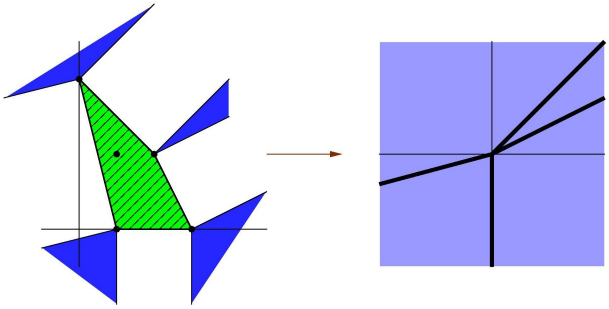

FIGURA 2: El politop de Newton de $f(\theta_1,\theta_2) = 2\theta_1^3 + 3\theta_1^2\theta_2^2 + \theta_1\theta_2^2 + 3\theta_1 + 5\theta_2^4$ i el seu ventall normal.

2 TEOREMA ([3]) *Siguin $P_1, P_2, \ldots, P_k$ politops en $\mathbb{R}^d$, i sigui m el nombre d'arestes no paral·leles de $P_1, \ldots, P_k$. Llavors, el nombre de vèrtexs de $P_1 + \cdots + P_k$ és com a molt*

$$2\sum_{j=0}^{d-1}\binom{m-1}{j}.$$

Observeu que aquesta fita és independent del nombre k de politops.

3 Una fita superior al nombre de funcions d'inferència

Per a paràmetres fixats, el problema d'inferència que consisteix a trobar l'explicació $\hat{\mathbf{h}}$ que maximitza $\text{Prob}(\mathbf{X} = \mathbf{h} \mid \mathbf{Y} = \tau)$ és equivalent a identificar el monomi $f_{\mathbf{h},\tau} = \theta_1^{a_{1,\mathbf{h}}}\theta_2^{a_{2,\mathbf{h}}}\cdots\theta_d^{a_{d,\mathbf{h}}}$ de f_τ amb valor més gran. Com que el logaritme és una funció monòtona creixent, el monomi desitjat també maximitza la quantitat

$$\begin{aligned}\log(\theta_1^{a_{1,\mathbf{h}}}\theta_2^{a_{2,\mathbf{h}}}\cdots\theta_d^{a_{d,\mathbf{h}}}) &= a_{1,\mathbf{h}}\log(\theta_1) + a_{2,\mathbf{h}}\log(\theta_2) + \cdots + a_{d,\mathbf{h}}\log(\theta_d)\\ &= a_{1,\mathbf{h}}v_1 + a_{2,\mathbf{h}}v_2 + \cdots + a_{d,\mathbf{h}}v_d,\end{aligned}$$

on hem substituït $\log(\theta_i)$ per v_i. Això és equivalent al fet que el punt corresponent $(a_{1,\mathbf{h}}, a_{2,\mathbf{h}}, \ldots, a_{d,\mathbf{h}})$ maximitza l'expressió lineal $v_1x_1 + \cdots + v_dx_d$ en el politop de Newton $\text{NP}(f_\tau)$. Per tant, el problema d'inferència per paràmetres fixos esdevé un problema de programació lineal.

Cada tria dels paràmetres $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ determina una funció d'inferència. Si $\mathbf{v} = (v_1, v_2, \ldots, v_d)$ és el vector d'$\mathbb{R}^d$ de coordenades $v_i = \log(\theta_i)$, denotem la funció d'inferència corresponent per

$$\Phi_{\mathbf{v}} : (\Sigma')^n \longrightarrow \Sigma^q.$$

Per a cada observació $\tau \in (\Sigma')^n$, la seva explicació $\Phi_{\mathbf{v}}(\tau)$ ve donada pel vèrtex de $\text{NP}(f_\tau)$ que és maximal en la direcció del vector $\mathbf{v}$. Noteu que per alguns

valors dels paràmetres (si $\mathbf{v}$ és perpendicular a una cara de dimensió positiva de $\mathrm{NP}(f_\tau)$) hi pot haver més d'un vèrtex que assoleixi el màxim. També és possible que un punt $(a_{1,\mathbf{h}}, a_{2,\mathbf{h}}, \ldots, a_{d,\mathbf{h}})$ en el politop correspongui a diversos valors diferents de les dades ocultes. En els dos casos triem l'explicació segons la regla per desfer empats determinada per endavant. Aquesta simplificació no afecta el nombre asimptòtic de funcions d'inferència.

Diferents valors de θ produeixen diferents direccions $\mathbf{v}$, que poden resultar en diferents funcions d'inferència. El nostre objectiu és fitar el nombre de funcions d'inferència diferents que un model gràfic pot tenir. El teorema següent dóna una fita superior que és polinòmica en la mida del model gràfic. De fet, molt poques de les $l^{q(l')^n}$ funcions $(\Sigma')^n \longrightarrow \Sigma^q$ són funcions d'inferència.

3 TEOREMA *Sigui d un enter positiu fixat. Considerem un model gràfic amb d paràmetres, i sigui E el nombre d'arestes del graf subjacent. Llavors, el nombre de funcions d'inferència del model és com a molt d'ordre $O(E^{d(d-1)})$.*

Abans de demostrar aquest teorema, observeu que el nombre E d'arestes depèn del nombre n de variables aleatòries observades. En quasi tots els models gràfics d'interès, E és una funció lineal de n, així que la fita esdevé $O(n^{d(d-1)})$. Per exemple, el model de Markov ocult té $E = 2n - 1$ arestes.

PROVA En la primera part de la prova reduirem el problema de comptar funcions d'inferència a l'enumeració dels vèrtexs d'un cert politop. Hem vist que una funció d'inferència queda especificada per l'elecció dels paràmetres, que és equivalent a triar un vector $\mathbf{v} \in \mathbb{R}^d$. La funció es denota per $\Phi_\mathbf{v} : (\Sigma')^n \longrightarrow \Sigma^q$, i l'explicació $\Phi_\mathbf{v}(\tau)$ d'una observació donada τ ve determinada pel vèrtex de $\mathrm{NP}(f_\tau)$ que és maximal en la direcció de $\mathbf{v}$. Per tant, els cons del ventall normal $\mathcal{N}(\mathrm{NP}(f_\tau))$ corresponen a conjunts de vectors $\mathbf{v}$ que produeixen la mateixa explicació per l'observació τ. Els cons no maximals (és a dir, els continguts en un altre con de dimensió més gran) corresponen a direccions $\mathbf{v}$ en les quals més d'un vèrtex és maximal. Com que desfem els empats amb una regla consistent, ignorarem aquest cas, per simplicitat. D'ara endavant considerarem només cons maximals del ventall normal.

Sigui $\mathbf{v}' = (v_1', v_2', \ldots, v_d')$ un altre vector, provinent d'una elecció diferent dels paràmetres (vegeu la figura 3). Pel raonament anterior, $\Phi_\mathbf{v}(\tau) = \Phi_{\mathbf{v}'}(\tau)$ si, i només si, $\mathbf{v}$ i $\mathbf{v}'$ pertanyen al mateix con de $\mathcal{N}(\mathrm{NP}(f_\tau))$. Per tant, $\Phi_\mathbf{v}$ i $\Phi_{\mathbf{v}'}$ són la mateixa funció d'inferència si, i només si, $\mathbf{v}$ i $\mathbf{v}'$ pertanyen al mateix con de $\mathcal{N}(\mathrm{NP}(f_\tau))$ per a totes les observacions $\tau \in (\Sigma')^n$. Considerem el refinament comú de tots aquests ventalls normals, $\bigwedge_{\tau\in(\Sigma')^n} \mathcal{N}(\mathrm{NP}(f_\tau))$. Llavors, $\Phi_\mathbf{v}$ i $\Phi_{\mathbf{v}'}$ són la mateixa funció d'inferència exactament quan $\mathbf{v}$ i $\mathbf{v}'$ pertanyen al mateix con d'aquest refinament comú. Això implica que el nombre de funcions d'inferència és igual al nombre de cons de $\bigwedge_{\tau\in(\Sigma')^n} \mathcal{N}(\mathrm{NP}(f_\tau))$. Pel lema 1, aquest refinament comú és el ventall normal de $\mathrm{NP}(\mathbf{f}) = \sum_{\tau\in(\Sigma')^n} \mathrm{NP}(f_\tau)$, la suma de Minkowski dels politops $\mathrm{NP}(f_\tau)$ per a totes les observacions τ. Així, doncs, enumerar les funcions d'inferència és equivalent a comptar els vèrtexs de $\mathrm{NP}(\mathbf{f})$. En la resta de la prova donarem una fita superior al nombre de vèrtexs de $\mathrm{NP}(\mathbf{f})$.

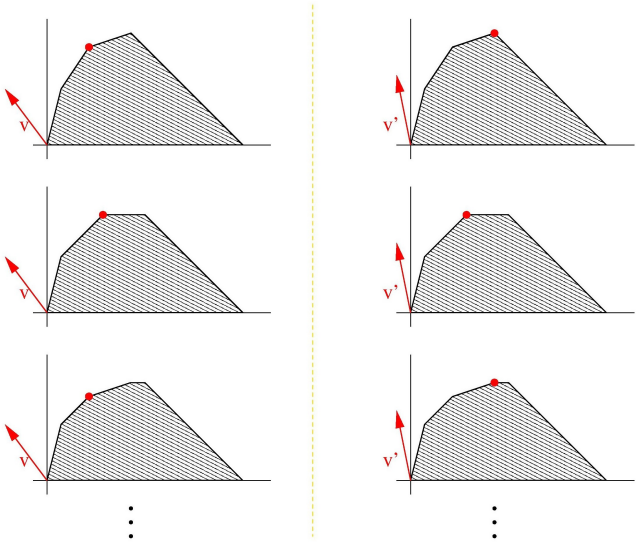

FIGURA 3: Dues funcions d'inferència distintes, $\Phi_{\mathbf{v}}$ (columna esquerra) i $\Phi_{\mathbf{v}'}$ (columna dreta). A cada fila hi ha el politop de Newton corresponent a una observació diferent. Les explicacions respectives vénen donades pels vèrtexs marcats en cada politop.

Observeu que, per cada τ, el politop $\mathrm{NP}(f_\tau)$ està contingut en l'hipercub $[0,E]^d$, ja que cada paràmetre θ_i pot aparèixer com a factor d'un monomi de f_τ com a màxim E vegades. A més, els vèrtexs de $\mathrm{NP}(f_\tau)$ tenen coordenades enteres, perquè són vectors d'exponents. Com a conseqüència, les arestes de $\mathrm{NP}(f_\tau)$ vénen donades per vectors on cada coordenada és un enter entre $-E$ i E. Només hi ha $(2E+1)^d$ vectors amb aquestes caraterístiques, i per tant $(2E+1)^d$ és una fita superior pel nombre de direccions diferents que les arestes dels politops $\mathrm{NP}(f_\tau)$ poden tenir.

Aquesta propietat dels politops de Newton de les coordenades del model ens permetrà donar una fita superior pel nombre de vèrtexs de la seva suma de Minkowski $\mathrm{NP}(\mathbf{f})$. L'últim ingredient que necessitem és el teorema 2. En el nostre cas tenim una suma de politops $\mathrm{NP}(f_\tau)$, un per a cada observació $\tau \in (\Sigma')^n$, de manera que el nombre total d'arestes no paral·leles és com a molt $(2E+1)^d$. Per tant, pel teorema 2, el nombre de vèrtexs de $\mathrm{NP}(\mathbf{f})$ és com a molt

$$2\sum_{j=0}^{d-1}\binom{(2E+1)^d-1}{j}.$$

Quan E tendeix a infinit, el terme dominant d'aquesta expressió és

$$\frac{2^{d^2-d+1}}{(d-1)!}E^{d(d-1)}.$$

Això ens dóna una fita superior $O(E^{d(d-1)})$ en el nombre de funcions d'inferència del model gràfic. □

A [2] es demostra que la fita donada pel teorema 3 és la millor possible, llevat d'un factor constant.

4 Funcions d'inferència per alineació de seqüències

En aquesta secció donem una aplicació del teorema 3 a un model bàsic per alineació de seqüències. L'alineació de seqüències s'utilitza per determinar la similitud entre seqüències biològiques que han evolucionat a partir d'un avantpassat comú a través d'una sèrie de mutacions, insercions i supressions. La resposta a quina és la millor alineació de dues seqüències donades depèn de la tria d'un esquema de puntuació. Per l'alineació paramètrica de seqüències es coneixen algorismes eficients (vegeu per exemple [7, capítol 7]). Aquí el que ens interessa són les diferents funcions d'inferència que poden sorgir quan els paràmetres varien. Per a un tractament detallat del tema d'alineació de seqüències podeu consultar [4].

Donades dues seqüències, σ^1 i σ^2, de longituds n_1 i n_2 respectivament, una *alineació* és una parella (μ^1, μ^2) de seqüències de la mateixa longitud que han estat obtingudes a partir de σ^1, σ^2 inserint-hi guionets $(-)$ de manera que no hi hagi cap posició on μ^1 i μ^2 tinguin un guionet a la vegada. Un *alineament* (*match*) és una posició on μ^1 i μ^2 tenen la mateixa lletra, un *no-alineament* (*mismatch*) és una posició on μ^1 i μ^2 tenen lletres diferents, i un *espai* és una posició on μ^1 o μ^2 tenen un guionet. Un sistema de puntuació senzill consisteix en dos paràmetres, α i β, que denoten les penalitzacions per a un *no-alineament* i per a un espai respectivament. La recompensa per a un *alineament* és 1. La puntuació d'una alineació amb z *alineaments*, x *no-alineaments* i y espais és doncs $z - x\alpha - y\beta$. Observeu que aquests nombres sempre satisfan $2z + 2x + y = n_1 + n_2$.

Aquest model per a l'alineació de seqüències és un cas particular de l'anomenat *model de Markov ocult aparellat*. El problema de determinar l'alineació amb màxima puntuació per a valors donats de α i β és equivalent al problema d'inferència en el model de Markov ocult aparellat. En aquest model, una observació és una parella de seqüències $\tau = (\sigma^1, \sigma^2)$, i el nombre de variables observades és $n = n_1 + n_2$. Els valors de les variables ocultes en una explicació indiquen les posicions dels espais en l'alineació òptima. Ens referirem a aquest model com el *model de* 2 *paràmetres per alineació de seqüències*.

Per cada parella τ de seqüències, el politop de Newton del polinomi f_τ és l'envolupant convexa dels punts (x, y, z) on les coordenades són el nombre de *no-alineaments*, espais i *alineaments*, respectivament, de cada possible alineació de la parella. Aquest politop és 2-dimensional, ja que pertany al pla $2z + 2x + y = n_1 + n_2$. En particular, no perdem informació si considerem la seva projecció en el pla xy. Aquesta projecció és simplement l'envolupant convexa dels punts (x, y) donats pel nombre de *no-alineaments* i espais de

cada alineació. Per a qualsevol alineació de seqüències de longituds n_1 i n_2, el punt corresponent (x, y) pertany al quadrat $[0, n]^2$, on $n = n_1 + n_2$. Per tant, com que tractem amb politops de coordenades enteres continguts a $[0, n]^2$, la prova del teorema 3 implica que el nombre de funcions d'inferència d'aquest model és $O(n^{2(2-1)}) = O(n^2)$. A continuació mostrem que aquesta fita quadràtica és ajustada, fins i tot en el cas de l'alfabet binari.

4 PROPOSICIÓ *Considerem el model de* 2 *paràmetres per alineació de seqüències, amb dues seqüències observades de longitud* n*, i sigui* $\Sigma' = \{0, 1\}$ *l'alfabet binari. Llavors, el nombre de funcions d'inferència d'aquest model és* $\Theta(n^2)$.

PROVA El raonament anterior prova que $O(n^2)$ és una fita superior pel nombre de funcions d'inferència del model. Per demostrar la proposició, hem de veure que el nombre d'aquestes funcions és com a mínim $\Omega(n^2)$.

Com que les dues seqüències tenen la mateixa longitud, el nombre d'espais en cada alineació és parell. Per comoditat, definim $y' = y/2$ i $\beta' = 2\beta$, i treballarem amb les coordenades (x, y', z) i els paràmetres α i β'. El valor y' s'anomena el nombre d'insercions (és la meitat del nombre d'espais), i β' és la penalització per una inserció. Per a valors fixos d'α i β', l'explicació d'una observació $\tau = (\sigma^1, \sigma^2)$ és donada pel vèrtex de $\mathrm{NP}(f_\tau)$, que és maximal en la direcció del vector $(-\alpha, -\beta', 1)$. En aquest model, $\mathrm{NP}(f_\tau)$ és l'envolupant convexa dels punts (x, y', z), on les coordenades són el nombre de *no-alineaments*, insercions i *alineaments* de les alineacions de σ^1 i σ^2.

El raonament de la demostració del teorema 3 prova que el nombre de funcions d'inferència d'aquest model és el nombre de cons del refinament comú dels ventalls normals de tots els politops $\mathrm{NP}(f_\tau)$, on τ és qualsevol parella de seqüències de longitud n en l'alfabet Σ'. Com que els politops $\mathrm{NP}(f_\tau)$ pertanyen al pla $x + y' + z = n$, és equivalent considerar els ventalls normals de les seves projeccions en el pla $y'z$. Aquestes projeccions són polígons amb coordenades enteres, continguts en el quadrat $[0, n]^2$. Denotarem per P_τ la projecció de $\mathrm{NP}(f_\tau)$ en el pla $y'z$.

A continuació construirem, per cada parell d'enters positius u i v primers entre si tals que $u < v$ i $6v - 2u \le n$, una parella $\tau = (\sigma^1, \sigma^2)$ de seqüències binàries de longitud n, de manera que P_τ tindrà una aresta amb pendent u/v. En el ventall normal $\mathcal{N}(P_\tau)$, aquesta aresta produeix la línia $u \cdot \alpha + v \cdot \beta' = 0$, que separa regions del ventall, i per tant la mateixa línia separadora també serà a $\bigwedge_\tau \mathcal{N}(P_\tau)$, on τ comprèn totes les parelles de seqüències binàries de longitud n. El nombre de possibilitats per u i v és $\Omega(n^2)$ (això es desprèn del fet que una proporció positiva dels possibles parells $(u, v) \in \mathbb{Z}^2$ satisfà que u i v són primers entre si; vegeu [1, capítol 3]). Per tant, el nombre de funcions d'inferència diferents és $\Omega(n^2)$.

Ara només ens queda construir τ amb les propietats indicades, donats dos enters positius u i v com els descrits abans. Siguin $a := 2v$ i $b := v - u$. Suposem primer que $n = 6v - 2u = 2a + 2b$. Considerem les seqüències

$$\sigma^1 = 0^a 1^b 0^b 1^a, \qquad \sigma^2 = 1^a 0^b 1^b 0^a,$$

on 0^a indica que el símbol 0 es repeteix a vegades. Sigui $\tau = (\sigma^1, \sigma^2)$. Llavors, es pot comprovar que el polígon P_τ per a aquesta parella de seqüències té quatre vèrtexs: $v_0 = (0,0)$, $v_1 = (b, 3b)$, $v_2 = (a+b, a+b)$ i $v_3 = (n, 0)$. El pendent de l'aresta entre v_1 i v_2 és $\frac{a-2b}{a} = \frac{u}{v}$.

Si $n > 6v - 2u = 2a + 2b$, simplement afegim $0^{n-2a-2b}$ al final de les dues seqüències σ^1 i σ^2. En aquest cas, els vèrtexs de P_τ són $(0, n-2a-2b)$, $(b, n-2a+b)$, $(a+b, n-a-b)$, $(n, 0)$ i $(n-2a-2b, 0)$.

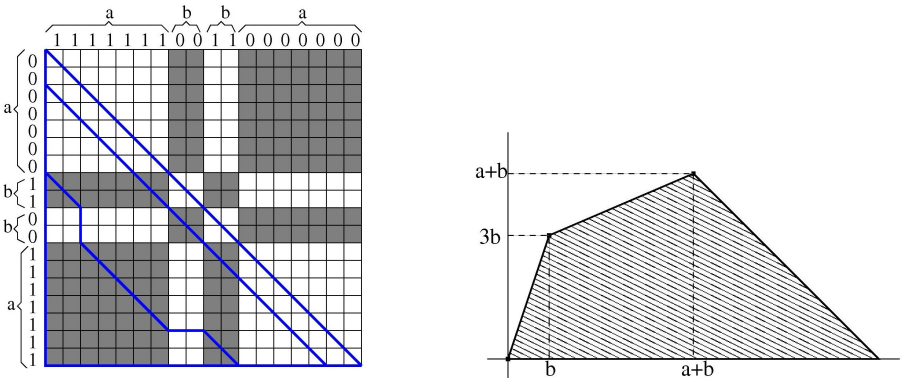

FIGURA 4: Parella de seqüències binàries de longitud 18 que produeix un pendent $3/7$ en el seu politop d'alineacions. Els quatre camins en el *graf d'alineacions* de l'esquerra corresponen als quatre vèrtexs; un pas a la dreta en el graf correspon a un espai en σ^1; un pas cap avall, a un espai en σ^2, i un pas diagonal, a un *alineament* o *no-alineament*. Podeu consultar [7, secció 2.2] per veure una definició completa del graf d'alineacions.

Observeu que, si $v - u$ és parell, la construcció es pot fer amb seqüències de longitud $n = 3v - u$ prenent $a := v$, $b := \frac{v-u}{2}$. La figura 4 mostra el graf d'alineacions i el polígon P_τ per $a = 7$, $b = 2$. □

En la majoria de casos, hom està interessat només en les funcions d'inferència que són significatives des d'un punt de vista biològic. Aquestes corresponen a valors dels paràmetres amb $\alpha, \beta \geq 0$, la qual cosa vol dir que els *no-alineaments* i espais són penalitzats en lloc de recompensats. A vegades també es requereix que $\alpha \leq \beta$, cosa que vol dir que un *no-alineament* és penalitzat menys de dos espais. La nostra construcció per a la prova de la proposició 4 demostra no solament que el nombre total de funcions d'inferència és $\Omega(n^2)$, sinó també que el nombre de les que són biològicament significatives és encara $\Omega(n^2)$.

5 Observacions finals

Una interpretació del teorema 3 és que la capacitat de canviar els valors dels paràmetres d'un model gràfic no dóna tanta llibertat com podria semblar. Hi ha un nombre molt gran de possibles maneres d'assignar una explicació a cada observació. Tanmateix, només n'hi ha una minúscula proporció que prové d'un mètode consistent a triar l'explicació més probable per a certs valors dels paràmetres. Tot i que els paràmetres poden variar de manera contínua, el nombre de funcions d'inferència diferents que es poden obtenir és polinòmic en el nombre d'arestes del graf, suposant que el nombre de paràmetres és fix.

Havent provat que el nombre de funcions d'inferència d'un model gràfic és polinòmic en la mida del model, el pas següent hauria de ser trobar una manera eficient de precomputar totes les funcions d'inferència per a models determinats. Això ens permetria donar la resposta (l'explicació) a una consulta (una observació) molt ràpidament. El teorema 3 suggereix que podria ser computacionalment factible precomputar el politop $\mathrm{NP}(\mathbf{f})$, els vèrtexs del qual corresponen a les funcions d'inferència. Tot i això, la dificultat sorgeix quan intentem descriure eficientment una funció d'inferència concreta. El problema és que la caracterització d'una funció d'inferència involucra un nombre exponencial d'observacions.

Referències

[1] Apostol, T. M. *Introduction to analytic number theory*. Nova York: Springer-Verlag, 1976.

[2] Elizalde, S.; Woods, K. «Bounds on the number of inference functions of a graphical model». [En preparació]

[3] Gritzmann, P.; Sturmfels, B. «Minkowski addition of polytopes: Computational complexity and applications to Gröbner bases». *SIAM Journal of Discrete Mathematics*, 6 (1993), 246–269.

[4] Gusfield, D. *Algorithms on strings, trees, and sequences*. Cambridge University Press, 1997.

[5] Gusfield, D.; Balasubramanian, K.; Naor, D. «Parametric optimization of sequence alignment». *Algorithmica*, 12 (1994), 312–326.

[6] Jensen, F. *Bayesian networks and decision graphs*. Springer, 2001.

[7] Pachter, L.; Sturmfels, B. [ed.] *Algebraic statistics for computational biology*. Cambridge University Press, 2005.

[8] Pachter, L.; Sturmfels, B. «The mathematics of phylogenomics». *SIAM Review*. [En premsa]

[9] Pachter, L.; Sturmfels, B. «Tropical geometry of statistical models». *Proc. Natl. Acad. Sci.*, 101, núm. 46 (2004), 16132–16137.

[10] Ziegler, G. M. *Lectures on polytopes*. Nova York: Springer, 1995. (Graduate Texts in Mathematics; 152).

Department of Mathematics
Dartmouth College
6188 Bradley Hall, Hanover, NH 03755
sergi.elizalde@dartmouth.edu